**분석 설계서**

1. **분석 배경 및 환경 정의**
2. 분석 배경
3. 기획 배경 : 사용자는 수많은 상품이나 정보 중 자신이 원하는 것을 편하게 찾고자 함. 따라서 쇼핑, 영화, 음악, 음식점과 같이 많은 분야에서 추천 시스템이 활용되고 있음. 하지만 기존의 추천 시스템은 사용자나 아이템의 유사도를 주천에 활용. 그러나 실제로 사용자가 대상을 선택 할 때는 현재 위치, 사용목적, 가격, 날씨, 시간 등 다양한 상황 요소를 고려하기 때문에 세밀한 분석이 필요.
4. 분석 환경 및 도구
5. 분석 환경 - Colab, Jupyter notebook, python, PyCharm
6. 분석 도구 - tensorflow, KoNLPy, NLTK Data\_stopword, Collections

1. **데이터 탐색(EDA, 전처리)**
2. 데이터 양: 레코드 수, 피처(변수) 수, 전체 용량 등
3. 서울 음식점 data (124500개) - id, 음식점명, 도로명/지번 주소, x좌표, y좌표, **리뷰 데이터**(네이버, 카카오, 망고플레이트, 다이닝코드, 구글), 메뉴, 영업시간
4. 피처 설명: 이름, 의미, 데이터타입, 결측값 개수 등
   1. 서울 음식점 리뷰 데이터
      1. 이름 : 's\_review’
      2. 의미 : 5개의 사이트(네이버, 카카오, 망고플레이트, 다이닝코드, 구글)에서 크롤링한 실제 리뷰 데이터들
      3. 타입 : string타입
      4. 결측값 개수 : 53942개
      5. 실 데이터 수 : 70558개

1. 데이터 정제 및 변환: 결측치, 이상치 처리, 파생변수 생성, 피처 추가 및 제거 등
   1. KoNLPy를 이용하여 토큰화
   2. 불용어 사전 적용
   3. 형태소 분석 수행(KoNLPy)
   4. 단어 빈도수 측정
   5. 불용어사전 최적화
2. 주요 피처 및 데이터 특징(비교, 관계, 분포, 추세) 파악: 기초 통계, 시각화 방법
   1. 벡터화 진행 후 빈도수 확인
3. **분석 피처 및 데이터 선정**
4. 분석을 위한 피처 선정
   1. 리뷰 데이터 : 서울 음식점 12만개를 5개의 사이트에서 크롤링하여 가져온 리뷰 데이터
5. 테스트용 데이터 선정 : 카카오맵 크롤링 데이터 10000개
6. **분석 모델 구현**
7. 샘플데이터 - 분석 모델 선정
   1. 형태소 분석 기법
      1. 자연어 문장으로 되어있는 음식점 평가 리뷰를 4개 사이트에서 수집.

목적 : 해당 데이터가 자연어로 구성된 다양한 평가의견을 포함함.

* + 1. 자연어 처리를 위해 문장의 단어 추출 기능을 제공하는 KoNLPy 라이브러리로 okt와 mecab을 활용. 조사, 어미, 문장 부호를 제외함.
    2. 이 중 특징을 나타낼 수 있는 단어만을 추출하기 위해 Stopwords를 사용한다. 이는 불용어 리스트로 여기에 포함된 단어를 제거한다.
    3. KoNLPy라이브러리 중 수행시간 및 결과의 정확도를 고려하고, 토큰화된 결과를 확인하여 okt로 결정
  1. 토픽모델링 기법
     1. 토큰화된 문장을 TF-IDF, LDA 토픽모델링, Word2Vec을 적용
     2. 리뷰별 keyword 추출을 확인
     3. 도출된 keyword를 word2vec으로 묶음
     4. pyLDAvis로 군집형성 점검
     5. 이후 군집별로 테마명 설정

1. 서비스에 활용한 종합 데이터
   1. 형태소 분석 기법
      1. 샘플데이터로 작성한 코드 활용
      2. stopwords 추가
      3. i번, ii번 과정 반복
   2. 토픽모델링 기법
      1. 토큰화된 문장을 LDA 토픽모델링, Word2Vec을 적용
      2. 리뷰별 keyword 추출을 확인
      3. 도출된 keyword를 묶음
      4. pyLDAvis로 군집형성 점검
      5. 주요 요소들을 바꾸며 모델 수정
      6. 토픽 할당
      7. 토픽별로 테마 설정